



関西学院大学リポジトリ

Kwansei Gakuin University Repository

「あなたの不満を買い取ります」の テキストマイニングに関する研究

著者	林 利憲
雑誌名	KGPS review : Kwansei Gakuin policy studies review
号	26
ページ	23-30
発行年	2019-03-31
URL	http://hdl.handle.net/10236/00027796

「あなたの不満を買い取ります」の テキストマイニングに関する研究

林 利憲*

【修士論文概要書】

【要旨】

本研究では、不満調査データセットに対してフリーソフトウェア KH Coder を用いてテキストマイニングを行った。ここでは、特にマーケティング分析、社会調査分析の二つの目的からの分析を行った。ここに、不満調査データセットとは、国立情報学研究所を通して株式会社 Insight Tech によって提供された不満投稿情報である。さらに、本研究では KH Coder 特有のコーディング機能を利用することで、コードに焦点を当てた分析を行った。前者のマーケティング分析を目的としたテキストマイニングでは、不満投稿情報の中でも企業への不満に該当する不満投稿の抽出を試み、企業間の比較分析を行った。その結果、各企業のポジションを視覚的に把握し、かつそれぞれのコードに関する企業間の統計的な有意差を示すことができた。得られた結果は、各企業に対する顧客の満足度向上のための提言や、競合相手との差別化を図るための提言を可能とする。社会調査を目的としたテキストマイニングでは、既婚女性による夫への不満、恋愛に対する不満の二つに絞り、投稿者の属性を利用したセグメンテーションを行った。その上で、各セグメントにおける不満傾向を導き出し、不満特性の差を明らかにした。

キーワード：テキストマイニング、不満調査データセット、KH Coder、マーケティング、社会調査

1. 緒言

近年、情報技術が凄まじい進歩を遂げたことにより、あらゆる分野で IT 化が進んでいる。世界中で多くの人間がスマートフォンと呼ばれる電子端末を持ち、様々な企業にも IT 化が浸透した。ネット上の決済やスマートフォンを用いた清算は当然となり、情報技術と生活の利便性は切り離せない社会となった。情報技術が進歩を遂げた一方、人々が発信するデータを収集・分析し、有益な情報を見つけ出すデータ分析に関する技術 [1, 2, 3, 4, 5, 6] も進歩している。

このような背景を基に本研究では、不満調査データセットを対象としたテキストマイニングを行った。ここに、不満調査データセットとは、株式会社 Insight Tech により公開された不満投稿プラットフォーム「あなたの不満を買い取ります」（通称：不満買取セン

* 関西学院大学大学院総合政策研究科博士課程前期課程 (den82687@kwansei.ac.jp)

ター)へ投稿された文書のデータセットであり、国立情報学研究所を通して、学術研究への利用を目的に国立情報学研究所に申請された多くの大学へ公開されている。不満投稿者は「あなたの不満を買い取ります」に無料登録をしたユーザで、投稿を重ねることでオンラインショッピングサイトに利用できるディスカウントチケットと交換することができる。株式会社 Insight Tech も収集したデータを分析し、企業に提供することで利益を得ている。本研究では、マーケティング分析・社会調査分析の二つを目的として、この不満調査データセットのテキストマイニングを行う。

2. 分析手法と分析データの構成

2.1 テキストマイニング手法

テキストマイニングには、より主体的、かつ明示的な分析を行うため、KH Coder [7]の中のコーディング機能を利用することとした。細かくルールを作成して文書を抽出、集計することによって、コード（コンセプト）により焦点を当てた分析を行うことが可能である。この KH Coder を利用することで、文書データからの特徴的な抽出語を集約したいいくつかのコードに対して、様々な統計的分析が可能となる。

本研究で用いたテキストマイニング手法は、(1) 前処理、(2) 統計的分析の 2 ステップから成る。特に、コーディングルールを用いて表現の統一化を図り、ルールベースで文書内の投稿抽出、及び統計的分析を行うことは従来のテキストマイニング分析とは異なる。(1) 前処理では「形態素解析」、「複合語の設定」、「不要語の設定」の三つを行う。形態素解析とは言語で意味を持つ最小単位である形態素毎に文章を分ける処理のことである。頻出語の抽出前に複合語と不要語を設定することで、以降の作業が容易となる。また、本研究では(2) 統計的分析において「コーディングルールの作成」、「対応分析（コレスポンデンス分析）」、「クロス集計分析」の三つを行うこととした。「コーディングルールの作成」では一つのコードと呼ばれるキーワードに複数の表現を集約するルールを作成する。これにより、似通った複数の表現を一つのコード（コンセプト）として統一された抽出が可能となり、分析が容易となる。なお、コーディングルールは論理演算子を用いた記述が可能である。この作成したコーディングルールを適用した集計結果を用いて「対応分析（コレスポンデンス分析）」による視覚的な分析、「クロス集計分析」による統計的有意差の検証を行う。

2.2 不満調査データセットの構成

本研究で取り扱う不満調査データセット [8]は、提供されたデータのうち、不満投稿情報、不満投稿者情報を利用した。提供を受けた不満投稿データは 2015 年の 3 月 18 日（サービス開始日）から 2017 年 3 月 12 日までに投稿された様々な不満に関する文書データであり、5,248,820 レコード、9 フィールドから構成されている。不満投稿者データは

182,242 レコード、9 フィールドから構成される。各フィールドには、投稿者の性別や生年、職業や所在都道府県まで格納されているので、投稿者をこうした属性に応じてグループに分けることにより、マーケティング分析で有用とされるセグメンテーション [9]を行うことができる。また、社会調査分析においても投稿者グループによる比較もできるので、特定の投稿者グループに焦点を当てた分析ができる。このような点から不満投稿者情報は豊富な属性情報が付随する利便性のあるデータといえる。

3. 企業比較によるマーケティング分析

マーケティング分析を目的としたテキストマイニングでは、不満投稿情報の中でも企業への不満に該当する不満投稿を抽出し、前述の統計的分析を行う。対象とする企業は（1）コンビニ企業 3 社、（2）情報通信企業 3 社、（3）e コマース企業 2 社、（4）居酒屋企業 2 社である。この結果、各企業のポジションを視覚的、かつ統計的な数値ともに示すことができ、顧客の満足度向上のためへの提言が可能となった。また、分析結果から競合相手との競争の上で、強みとなるような要素についても同様である。本稿では、（1）のコンビニ企業 3 社の分析結果について説明し、その他の企業間比較は分析結果を要約する。

3.1 コンビニ企業 3 社比較

コンビニ企業 A、B、C の 3 社に関する分析では、「接客態度」、「価格表示」、「ポイントカード」、「価格設定」、「品揃え」、「コーヒー」、「ドーナツ」の七つのコードを作成し、3 社間を比較する。図 1 にコンビニ企業比較のために作成したコーディングルールを示す。コーディングルールの詳しい作成方法は修士論文を参照されたい。

図 2 にコーディングルールを適用した対応分析の結果を示す。この結果から、A 社には「ドーナツ」、B 社には「接客態度」、C 社には「ポイントカード」との結びつきが強いことが分かった。また、表 1 はコーディングルールを適用したクロス集計結果を示す。この結果から、次の仮説を用いて分割表による独立性の検定 [10, 11, 12]を行う。

■ H_0 （帰無仮説）：外部変数とコードは独立である

■ H_1 （対立仮説）：外部変数とコードは独立でない

検定の結果、「接客態度」、「ポイントカード」、「価格設定」、「品揃え」、「ドーナツ」の五つのコードのそれぞれと 3 社の企業は独立でない、すなわち 3 社間に統計的有意差が認められた。特に「接客態度」に関する不満割合は全体の中でも大きいので、「接客態度」の不満割合が最も大きい B 社は改善する必要があるといえる。

```

コンビニBig3_コーディング1216.txt * - TeraPad
ファイル(F) 編集(E) 検索(S) 表示(V) ウィンドウ(W) ツール(T) ヘルプ(H)

1 *接客態度↓
2 接客 | 態度 | 店員 | タメ | 不愛想 | 対応 | 表情↓
3 ↓
4 *価格表示↓
5 ( 価格 | 値段 ) & 表示↓
6 ↓
7 *ポイントカード↓
8 ポイント | Point↓
9 ↓
10 *価格設定↓
11 ( 価格 | 値段 | 料金 | '料' ) | ( 高い | たかい | 安い | やすい ) ↓
12 ↓
13 *品揃え↓
14 売り切れ | '品揃え' | ( 品 & 少ない ) | Sold | 品切れ | 品薄 | 完売 | ( 種類 &
15 ( 少ない | '増やして' ) ) | 在庫 | 欠品↓
16 ↓
17 *コーヒー↓
18 コーヒー | ミルク | 'カフェ' | 珈琲↓
19 ↓
20 *ドーナツ↓
21 ドーナツ | ドーナッツ | ナッツ↓
22 [EOF]

22行: 1桁 標準 [80] SJIS CRLF 挿入

```

図 1 コンビニ企業3社分析用コーディングルール

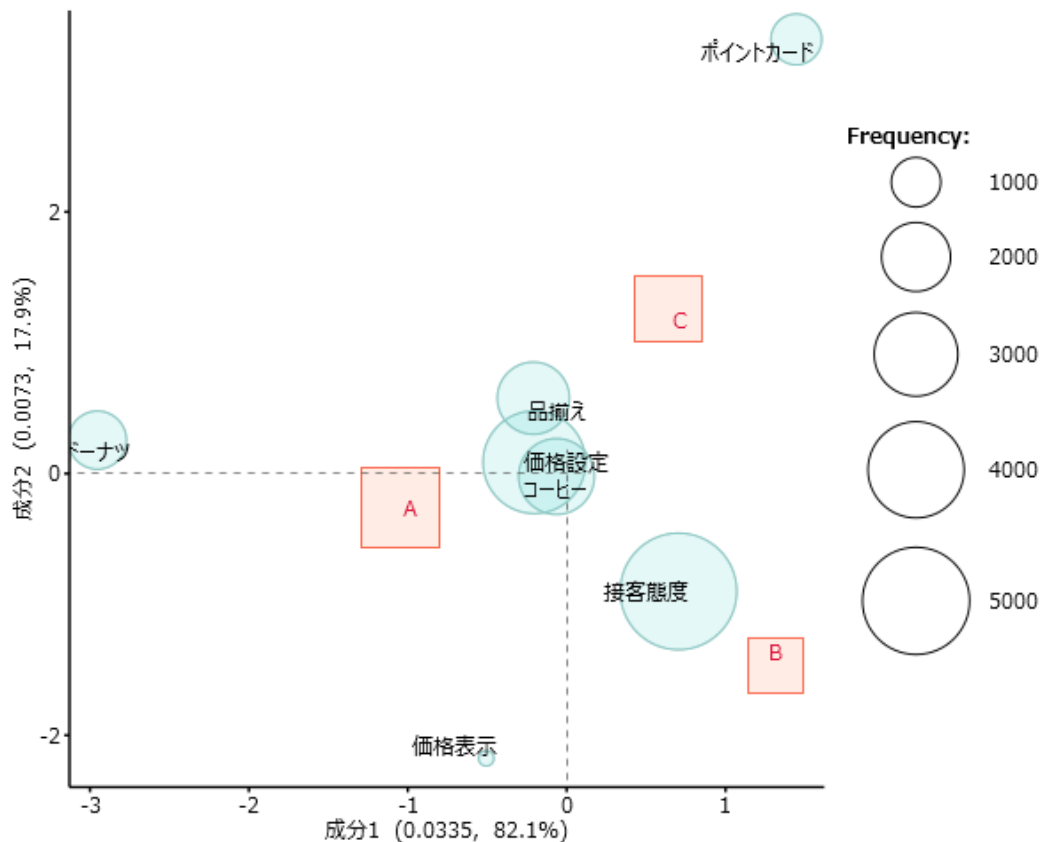


図 2 コンビニ主要3社を比較したクロス集計分析結果

表1 コンビニ主要3社を比較したクロス分析結果

	*接客 態度	*価格表示	*ポイント カード	*価格 設定	*品揃え	*コーヒー	*ドーナツ	ケース数
A社	2434 (11.53%)	36 (0.17%)	319 (1.51%)	2194 (10.39%)	1048 (4.96%)	1145 (5.42%)	1011 (4.79%)	21,110
B社	1561 (15.45%)	16 (0.16%)	203 (2.01%)	880 (8.71%)	396 (3.92%)	494 (4.89%)	73 (0.72%)	10,102
C社	1954 (13.20%)	16 (0.11%)	541 (3.66%)	1503 (10.15%)	763 (5.16%)	807 (5.45%)	318 (2.15%)	14,801
合計	5949 (12.93%)	68 (0.15%)	1063 (2.31%)	4577 (9.95%)	2207 (4.80%)	2446 (5.32%)	1402 (3.05%)	46,013
カイ2乗値	94.816**	2.397	182.405**	22.629**	22.466**	4.676	442.098**	

3.2 情報通信企業3社比較

情報通信企業3社の比較では、D社、E社、F社の3社間比較を行った。導入したコードは、「料金」、「契約内容」、「機種変更手続き」、「接客サービス」、「電波状態」、「通信速度」、そして「宣伝広告」の七つである。コーディングルールを適用した対応分析の結果、対象とした3社と各コードの関連性を明らかにした。クロス集計分析における分割表による独立性の検定では、「接客サービス」、「通信速度」、「宣伝広告」のコードで、3社間に統計的な有意差が認められた。特に「接客サービス」に関する不満割合は大きかったため、3社で最も不満割合が大きいE社は率先して解消すべき不満であることが示唆された。

3.3 eコマース企業2社比較

eコマース企業2社の比較では、G社、H社の2社を「商品の梱包」、「キャンペーン」、「価格設定」、「サイト内表示」、「品揃え」、「広告宣伝」、「アプリ」の七つのコードを用いて比較した。コーディングルールを適用した対応分析の結果から、G社には「商品の梱包」、H社には「キャンペーン」、「広告宣伝」のコードとの結びつきが強いことが分かった。また、クロス集計分析で行った分割表による独立性の検定では、「商品の梱包」、「キャンペーン」、「価格設定」、「サイト内表示」、「広告宣伝」、「アプリ」のコードに該当する不満割合で、2社間に統計的な差が認められた。特に「価格設定」、「サイト内表示」のコードに該当する不満は、全体の15%以上を占めるほど大きな割合のため、「価格設定」の不満割合が特に大きいG社、「サイト内表示」の不満割合が特に大きいH社には改善が求められる。

3.4 居酒屋企業 2 社比較

居酒屋を営む企業 2 社の比較では、大手居酒屋企業 2 社（I 社、J 社）を「食べ物の質」、「メニュー」「提供までの時間」、「お酒」、「価格」、「環境」、「サービス」のコードを用いて比較した。対応分析の結果、J 社には「食べ物の質」、「環境」、I 社には「価格」のコードとの間に相対的に強い関連があることが示された。分割表による独立性の検定の結果からは「食べ物の質」、「価格」、「環境」、「サービス」のコードで統計的に有意差が認められた。特に I 社は「価格」、J 社には「食べ物の質」、「環境」、「サービス」に関する不満割合が大きいため、それぞれに改善の必要があると考えられる。

4. ユーザ属性間比較による社会調査分析

社会調査を目的としたテキストマイニングでは、（5）家族・親戚サブカテゴリに該当する不満の中から既婚女性による夫への不満、そして（6）恋愛サブカテゴリに該当する不満の中から恋愛に対する不満の2種類に絞り、ユーザの不満傾向の分析を行う。夫への不満に対しては、投稿者の子供数・年収による属性から投稿者をグループに分け（セグメンテーション）、不満の傾向を導き出す。さらに、恋愛に対する不満分析では性別による恋愛観の違い・所在地域による不満特性の違いを明らかにする。

4.1 夫への不満分析

既婚女性による夫への不満の分析では、「配偶者の親」、「家事」、「育児」、「日常習慣」、「金銭」の五つのコードを作成した。投稿者の子供数による比較の結果、対応分析ではコードと投稿者の子供数の関連性を視覚的に明示することができた。また、クロス集計分析の結果、「配偶者の親」、「家事」、「育児」、「金銭」のコードで統計的に有意差が認められた。特に「配偶者の親」、つまり姑や舅に関する不満は投稿者の子供数が増えるにつれて、不満割合が減るという結果となった。また、投稿者の年収による比較では投稿者の年収を大・中・小に分けて比較した。対応分析の結果、各コードと投稿者の年収との関連性を明らかにし、クロス集計分析の結果より「配偶者の親」、「育児」、「日常習慣」、「金銭」のコードに該当する不満割合で投稿者を子供の人数に応じてグループ分けした場合、グループ間に統計的有意差が認められた。これらのコードに対して、年収が大きくなるにつれて不満割合が小さくなるという傾向が示された。

4.2 恋愛に対する不満分析

恋愛に対する不満分析には「外見」、「性格」、「連絡」、「他の人と遊び」、「出会い」、「遠距離恋愛」、「同性愛」の七つのコードを採用し、投稿者の性別・所在地域による分析を行った。コーディングルールを適用した対応分析の結果から、女性には「元恋

人」、「恋愛」に関する不満、男性には「同性愛」に関する不満との関連があることが示された。分割表による独立性の検定の結果、「元恋人」、「連絡」、「他の人と遊び」、「出会い」、「同性愛」の五つのコードで、投稿者の性別間に統計的有意差が認められた。これらの結果から、性別間の恋愛に対する不満の傾向を明らかにすることができた。一方、投稿者の所在地域による比較では、外部変数として取り込んだ地域の数が多いため、対応分析の結果からは明らかな特徴は得られなかった。しかし、クロス集計分析の結果から、コード「他の人と遊び」に該当する不満割合において地域間に統計的有意差が認められた。さらに、近畿地方より南の地方では、このコードに該当する不満割合が大きくなるという特徴を明らかにすることができた。

5. コードを適用した分析自動化の検討

以上に展開したマーケティング分析、社会調査に関する分析結果に加え、本研究ではコーディングルールの作成の自動化についても検討を加えた。そこでは、トピックモデルとも呼ばれる LDA (Latent Dirichlet Allocation) [13]を用いたシステムのみでの処理によるコードを適用した分析の代用を試みた。その結果、LDA に限界があり、他手法の応用の検討が必要であることを確かめた。

6. 結言

本研究では、不満調査データセットを対象としてマーケティング分析、社会調査分析を目的にフリーソフトウェアKH Coderを用いてテキストマイニングを行い、カテゴリ毎に分析、考察を行った。マーケティング分析を目的としたテキストマイニングでは、不満投稿情報の中でも企業への不満に該当する不満投稿を抽出し、分析を行った。分析の結果、各企業のポジションを視覚的、かつ統計的な数値ともに示すことができ、それらの結果を用いて、顧客の満足度向上のための提言が可能となった。また、分析結果から競合相手との競争で、強みとなるような要素の提言も行った。社会調査を目的としたテキストマイニングでは、既婚女性による夫への不満、そして恋愛に対する不満に絞り、ユーザの不満傾向の分析を行った。投稿者の属性を用いたさらなるセグメンテーションを行ったマーケティング分析、幅広いカテゴリに該当する不満のテキストマイニング、加えてコーディングルールの代用に向けた様々な手法の検討も必要であるが、今後の課題としたい。

謝辞

本研究を進めるにあたり、研究をする機会を失い、途方に暮れていた私を拾っていただき、終始熱心なご指導を頂いた三道弘明教授に深い謝意を表します。また、お忙しいところ、副査を引き受けてくださり、本論文作成の上で有益な助言をしてくださった山田孝子教授、伊佐田百合子教授に感謝申し上げます。精神的に追い詰められた時等、研究に関わらず様々なことに相談にのってくださった荻野哲男助手に厚く感謝致します。常に切磋琢磨できる仲間であった成川健太郎さんにも、刺激をいただき、精神的に支えられました。心から感謝致します。

本研究では、株式会社 Insight Tech が国立情報学研究所の協力により研究目的で提供している「不満調査データセット」を利用しました。データのご提供、感謝致します。

【参考文献】

- [1] 三道弘明, “楽天トラベル評価データのテキストマイニング,” 日本マーケティング・サイエンス学会, Vol.25, No.1, pp.62-63, 2018.
- [2] 堀部めぐみ, 笹岡沙也加, 長沼美沙, 長谷川栞, 原英彰, 中村光浩, “テキストマイニングによる産後うつについて母親が思うことの分析,” 看護科学研究 Vol.16, No.2, pp.53-63, 2018.
- [3] 小木しのぶ, “テキストマイニングの技術と動向,” 計算機統計学, Vol.25, No.1, pp.31-40, 2015.
- [4] 田口哲也, 野中誠, “不動産オンラインレビューを利用した住環境ニーズの分析,” 経営情報学会 2018 年春季全国研究発表大会要旨集, pp.210-213, 2018.
- [5] 金明哲, “教育と研究のためのテキストマイニングツール MTMineR (5.4),” 日本計算機統計学会大会論文集, Vol.30, pp.113-116, 2016.
- [6] 阿濱志保理, “質的分析に基づいた知的財産に対する学習意識の解明,” 産学連携学, Vol.14, No.2, pp.31-40, 2018.
- [7] 樋口耕一, “テキスト型データの計量的分析ー2つのアプローチの峻別と統合ー,” 理論と方法, Vol. 19, No.1, pp.101-115, 2004.
- [8] K. Mitsuzawa, M. Tauchi, M. Domoulin, M. Nakashima and T. Mizumoto, "FKC Corpus: a Japanese Corpus from New Opinion Survey Service," Proceedings of the Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results, pp.11-18, 2016.
- [9] P. Kotler, K. L. Keller, 恩藏直人 (監) and 月谷真紀 (訳), コトラー&ケラーのマーケティング・マネジメント 第12版, 丸善出版, 2014.
- [10] Frank Yates, “Contingency tables involving small numbers and the χ Test,” Supplement to the Journal of the Royal Statistical Society, Vol.1, No.2, pp.217-235, 1934.
- [11] 国沢清典 (編), 確率統計演習 2, 培風館, 1996.
- [12] 薩摩順吉, 確率・統計 (理工系の数学入門コース 7), 岩波書店, 1989.
- [13] D. M. Blei, Y. A. Ng and I. M. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, Vol.3, pp.993-1022, 2003.